DOCUMENT RESUME

ED 161 896 TM 007 423

AUTHOR Sachar, Jane; Suppes, Patrick
TITLE Estimating Total-test Scores from Partial Scores in a Matrix Sampling Design.
PUB DATE [77]
NOTE 22p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Comparative Statistics; Elementary Education; Item Analysis; Item Banks; *Item Sampling; Reliability; *Scores; *Testing
IDENTIFIERS Multiple Matrix Sampling

ABSTRACT

It is sometimes desirable to obtain an estimated total-test score for an individual who was administered only a subset of the items in a total test. The present study compared six methods, two of which utilize the content structure of items, to estimate total-test scores using 450 students in grades 3-5 and 60 items of the 110-item Stanford Mental Arithmetic Test. The items were sampled in such a way as to make comparisons between overlapping subtest designs and nonoverlapping subtest designs feasible. Three methods yielded fairly good estimates of the total-test score, namely regression with perfectly correlated nonoverlapping item samples, regression with correlation between item samples on overlapping subtests, and perfectly parallel overlapping or nonoverlapping item samples. The second method is suggested to be more robust than the other two and is, therefore, recommended. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Jane Sachar

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

ED161896

TM007423

Estimating Total-test Scores from Partial Scores in a Matrix Sampling Design

Jane Sachar[1]

Navy Personnel Research and Development Center

San Diego, California

Patrick Suppes

Institute for Mathematical Studies in the Social Sciences

Stanford University

Running head: Estimating Total-test Scores

[1]The views expressed herein are those of the author and do not necessarily reflect official Navy policy. This work was performed under contract number AID/CM-ta-C-73-40.

ABSTRACT

In establishing national test norms, sampling both examinees and items serves to reduce the amount of testing time required. It is often desired to obtain a total-test score for an individual who was administered only a subset of the total test. The present study compared six methods, two of which utilize the content structure of items, to estimate total-test scores using 450 students and 60 items of the 110-item Stanford Mental Arithmetic Test. The items were sampled in such a way as to make comparisons between overlapping subtest designs and nonoverlapping subtest designs feasible. Three methods yielded fairly good estimates of the total-test score, namely regression with perfectly correlated nonoverlapping item samples, regression with correlation between item samples on overlapping subtests, and perfectly parallel overlapping or nonoverlapping item samples. The second method is suggested to be more robust than the other two and is, therefore, recommended.

Estimating Total-test Scores from Partial Scores in a Matrix Sampling Design

Matrix sampling has been demonstrated to be a useful method for establishing national test norms (Lord, 1962). Sampling both examinees and items serves to reduce the amount of testing time required of each examinee. Analysis of resultant test data is based upon the assumption that the sample of items and the sample of examinees are drawn independently and that responses to an item do not depend on the context in which the item is presented.

Various investigators have attempted to estimate parameters of the total-test distribution of all examinees, even though each examinee receives only a portion of the items on the total test. Most studies consist primarily of estimates of the total-test mean and variance (Owens and Stufflebeam, 1969; Plumlee, 1964). Sirotnik and Wellington (1977) also provide methods for estimating means and variance components using an analysis of variance framework. Five studies included estimates of the total-test distribution (Cook and Stufflebeam, 1967 a, b; Lord, 1962, Kleinke, 1972; Bunda, 1973; Jaeger, in press). The distribution is essential to the estimation of percentile rankings. Lord and Cook and Stufflebeam fitted a negative hypergeometric distribution to three parameters, namely an estimated mean, an estimated variance, and the number of items on the total test.

It is sometimes the case that information is desired on the performance of a group of subjects on a large set of items and that information on the performance of the individual is also needed for purposes of individual evaluation. A matrix sampling design is ideal for the first purpose, but a method for estimating individuals' scores is essential to individual evaluation.

With the exception of Kleinke, Bunda, and Jaeger, none of the authors estimated total scores for individuals who were administered partial tests. Bunda estimated total scores from overlapping item samples using a regression equation whose coefficients are found from the item-total covariance matrix, the item variance-covariance matrix, and the item means.

Kleinke offered a method for nonoverlapping item samples using a linear prediction approach for generating the estimated total-test distribution. With this approach, the total test may be considered a composite of two tests, X, consisting of the items presented to the student, and Y, consisting of the items not presented to the student. The observed score on X is used to predict the score on Y. The predicted total-test score is then the sum $X + \hat{Y}$. For student i, the prediction of Y, namely $\hat{Y}_i$, is found by

$$\hat{Y}_i = c(X_i - \bar{X}) + \bar{Y}$$

where X_i is his score on X, $\bar{X}$ is the mean of the scores obtained on X, $\bar{Y}$ is the estimated mean on Y, and c is an unknown constant.

Jaeger offered two additional approaches of item sampling and estimation of total scores, referred to as "simple random sampling" and "stratified random sampling." Using simple random sampling, items are sampled from the total test with equal probabilities and without replacement in such a way as to produce a number of item subsets equivalent to that of a balanced incomplete block design. Using stratified random sampling, items are first stratified by some relevant criterion and then the above procedure is applied. From the score on the subset of items taken, X, the total test score, $\hat{Y}$, is estimated as

$$\hat{Y}_i = \frac{K}{k} X_i$$

where K is the number of items on the total test and k is the number of items on the subset of items taken by the student. With simple random sampling, the estimate is unbiased. With stratified random sampling, it is unbiased only when item samples drawn from each stratum are precisely proportional to the sizes of the strata on the total test.

Rasch (1960) developed a model which may be used to predict total-test scores from a matrix sampling design, although it was not designed for such a purpose. Using this model, one may calculate two sets of parameters, "easiness" parameters for items and "competency" parameters for individuals. An important consequence of this model is that the number of correct responses to a given set of items is a sufficient statistic for estimating a person's competency (Wright and Panchapakesan, 1968). Thus, from a set of competency tables calibrated through the inclusion of linking items (with overlapping tests), total-test scores can be predicted.

In summary, there are a number of alternative approaches for predicting total-test scores while using multiple matrix sampling. We present a matrix sampling design, three of the above methods, three new methods, and results of an empirical analysis which allowed for a comparison of overlapping and non-overlapping designs. Two of the new methods utilize content of test items, as well as statistical relationships, to provide score estimates.

Theoretical Basis

We will assume that examinees are randomly divided into v groups ($v = t^2$ by notation of Shoemaker and Knapp, 1974) and each group is adminstered a subtest of k items from a total test containing K items. We seek a method for

estimating each examinee's total-test score on the K items. To compare approaches using both overlapping and nonoverlapping item samples, we have structured the test to be analyzed in two ways. Each student is administered one subtest, containing one type-A item sample and one type-B item sample, as shown in Figure 1. The total test consists of K(a) type-A items and K(b) type-B items[K(a) + K(b) = K]. There is no overlap within type, i.e., no student is administered items from two different type-A item samples. Thus, methods for predicting total-test scores using Nonoverlapping Multiple Matrix Sampling (NMMS) designs may be applied to the K(a) type-A items independently of the K(b) type-B items. The total-test score for the K items is merely the sum of the K(a) and K(b) items. However, there is overlap across type, i.e., students taking different subtests may have the same item sample of type A or type B (but not both). Thus, methods for predicting total-test scores using overlapping item samples may also be applied. Presented below are six methods for estimating total-test scores. Two of the methods discussed above could not be employed on the matrix sampling design in this study: Bunda's model requires that every pair of items be administered to some students to obtain the inter-item covariance matrix and Rasch's model requires that all items fit the model in order to calibrate items, which the items employed here do not.

Method 1: Perfectly correlated nonoverlapping item samples (Kleinke's model with r = 1). Let the following be represented within item type:

X_i = the score on the item sample student i was administered

$\bar{X}$ = the mean on that item sample, taken across students

Sx = the standard deviation on that item sample

$\bar{Y}$ = the sum of the means on all other nonoverlapping item samples of the same type.

Sy = the estimated standard deviation of the sum of those item samples

r_{xy} = the correlation between item sample X and the sum of the other item samples, Y

Then student i's estimated score on the item samples, Y, using the standard regression equation, is given by:

$$\hat{Y}_i = r_{xy} \frac{Sy}{Sx} (X_i - \bar{X}) + \bar{Y}.$$

Method 1 assumes that there is a perfect correlation between test X and test Y, i.e., $r_{xy} = 1$, and therefore,

$$\hat{Y}_i = \frac{Sy}{Sx} (X_i - \bar{X}) + \bar{Y}.$$

The predicted total test score is the sum of X_i and $\hat{Y}_i$ for type-A items and X_i and $\hat{Y}_i$ for type-B items.

Method 2: Zero correlation between nonoverlapping item samples. (Kleinke's model with r = 0). This method assumes there is no correlation between test X and test Y, i.e., $r_{xy} = 0$; in this case

$$\hat{Y}_i = \bar{Y}.$$

Method 3: Correlation between item samples on overlapping subtests. This method applies a different regression model from that of Methods 1 and 2. In the general model it is more likely that $0 < r_{xy} < 1$. Method 3 utilizes overlapping item samples to estimate the coefficients in a regression equation on two item samples. Briefly, the method involves finding the correlation between

With an appropriate matrix design, taking all groups who took the same type-A item sample that examinee sample G took, the appropriate regression equation yields estimated scores for each of the type-B item samples that a student in examinee sample G did not take. The same procedure matching type-B item samples yields estimates for all type-A item samples. The predicted total score is then the sum of the two observed item sample scores and the estimated scores from the item samples not taken.

Method 4: Perfectly parallel overlapping or nonoverlapping item samples (Jaeger's model--equal means and standard deviations, r = 1). If the items are randomly assigned to subtests by simple random sampling, stratified random sampling, or systematic sampling, and the subtests are randomly assigned to examinees, then an examinee's score on any subtest may be estimated by his score on the subtest that was administered to him. To obtain the total score, Jaeger's method multiplies the examinees score by a constant reflecting the proportion of items from the total test which were administered to the subject. If there are K items on the total test and the subject gets the score X_i on both subtests consisting of 2k of these items, then his predicted total score is

$$\hat{Y}_i = \frac{K}{2k} X_i.$$

The predicted total score with this model is simply $\hat{Y}_i$.

Method 5: Item content with uncorrelated structural variable coefficients on overlapping or nonoverlapping subtests. Methods 1 through 4 and the previous studies cited provide score estimates that are independent of item content. Methods 5 and 6 of the present study use information about the structure of the individual items to predict student scores. These structural variables

are not those used in path analysis, but instead describe the characteristics of the items in terms of their content. The methodology applies when the items contained in the test can be described by structural variables. As an example, the structural variables on an arithmetic test may be (a) type of operation, (b) mode of presentation (oral or written), and (c) the number of steps required. The probability correct for an item can be found as a linear combination of these structural variables. (The use of structural variables for analyzing difficulty of arithmetic exercises is described in detail in Suppes and Morningstar, 1970).

Let (x_i), $i = 1,..,m$ be the structural variables and (a_i), $i = 1,..,m$ be random variables with $a_i \sim N(u_i, s_i^2)$. The general model for the probability correct, p, employs a transformation on p, namely $p' = \log p/(1 - p)$.

Then

$$p' = \sum_i a_i x_i + e$$

where the errors, e, are identical and independently distributed with a mean of zero, and variance s, and are independent of the a_i's. The expected value of p' is

$$E(p') = E\left[\sum_i (a_i x_i + e)\right] = \sum_i E(a_i)x_i = \sum_i u_i x_i$$

and the variance of p' is

$$V(p') = E\left[\sum_i (a_i x_i + e) - E\left(\sum_i a_i x_i + e\right)\right]^2$$

$$= \sum_i s_i^2 x_i^2 + 2 \sum_{i<j} s_i s_j x_i x_j r_{ij} + s_e^2$$

where r_{ij} is the correlation between variables a_i and a_j.

We can estimate u_i and s_i as follows: Randomly partition the sample into g subgroups (g chosen so that subgroups have adequate size). By solving the appropriate normal equations we are provided with g estimations of the mean, u_i, of each random variable a_i, i = 1,..,m. This allows for estimates of u_i and s_i as follows. For the description of method and for later use, we take g = 7 and obtain

$$p'^{j} = a_1^j x_1 + a_2^j x_2 + \ldots + a_m^j x_m + e^j \quad j = 1,\ldots,7.$$

As estimates, we take

$$\hat{u}_i = \frac{1}{7}\sum_j (u_i^j)$$

and

$$\hat{sm}_i^2 = \frac{1}{6}\sum_j (a_i^j - \hat{u}_i)^2,$$

where $\hat{sm}_i^2$ is an estimate of the variance of the mean, which is equal to s_i^2/N. Therefore, an estimate of the variance of a_i is simply $N\,\hat{sm}_i^2$, where N is the number of subjects in the sample. For each item, t, we write

$$p'_t = \sum_i a_i x_{it} + e_t$$

$$E(p'_t) = \sum_i u_i x_{it}$$

$$V(p'_t) = \sum_i s_i^2 x_{it}^2 + 2\sum_{i<j} s_i s_j x_{it} x_{jt} r_{ij} + s_e^2$$

Proceeding, the total scores will be

$$Y = \sum_t p_t' = \sum_t \sum_i (a_i x_{it} + e_t)$$

$$E(Y) = \sum_t E(p') = \sum_t \sum_i u_i x_{it}$$

$$V(Y) = \sum_t \left(\sum_i^t s_i^2 x_{it}^2 + 2 \sum_{i<j} s_i s_j x_{it} x_{jt} r_{ij} + s_e^2 \right)$$

Thus, the total score has the following distribution:

$$Y \sim N \left\{ \left(\sum_t \sum_i u_i x_{it} \right), \sum_t \left(\sum_i s_i^2 x_{it}^2 + 2 \sum_{i<j} s_i s_j x_{it} x_{jt} r_{ij} + s_e^2 \right) \right\}$$

Let Z_p be the standard score or z-score for student p,

$$Z_p = \frac{Y_p - E(Y)}{V(Y)^{1/2}}$$

Then the total score for student p will be

$$\hat{Y}_p = E(Y) + Z_p V(Y)^{1/2}$$

$$= \sum_t \sum_i u_i x_{it} + Z_p \left[\sum_t \left(\sum_i s_i^2 x_{it}^2 + 2 \sum_{i<j} s_i s_j x_{it} x_{jt} r_{ij} + s_e^2 \right) \right]^{1/2}$$

The x_{it}'s are structural variables and the u_i's and s_i's can be estimated as above. s_e^2 is estimated using the errors from the model including p's for all subjects on all items taken. The correlation, r_{ij}, between the structural variables a_i and a_j is unknown. However, the upper and lower bounds on the student's score can be determined by predicting the score assuming perfect correlation ($r_{ij} = 1$) and by predicting the score assuming no correlation ($r_{ij} = 0$). For uncorrelated structural variables, we set $r_{ij} = 0$ for all pairs i,j. The variance becomes

$$V(Y) = \sum_t \left(\sum_i s_i^2 x_{it}^2 + s_e^2 \right)$$

and the predicted score becomes

$$\hat{Y}_p = \sum_t \sum_i u_i x_{it} + Z_p \left[\sum_t \left(\sum_i s_i^2 x_{it}^2 + s_e^2 \right) \right]^{1/2}$$

Method 6: Item content with perfectly correlated structural variable coefficients on overlapping or nonoverlapping subtests. For perfectly correlated structural variables, we set $r_{ij} = 1$ for all pairs i,j. The variance becomes

$$V(Y) = \sum_t \left(\sum_i s_i^2 \, x_{it}^2 + 2 \sum_{i<j} s_i s_j x_{it} x_{jt} + s_e^2 \right)$$

$$= \sum_t \left[\left(\sum_i s_i x_{it} \right)^2 + s_e^2 \right]$$

and the predicted score becomes

$$\hat{Y}_p = \sum_t \sum_i u_i x_{it} + Z_p \left\{ \sum_t \left[\left(\sum_i s_i x_{it} \right)^2 + s_e^2 \right] \right\}^{1/2}$$

Empirical Comparison

Approach

We demonstrate these six methods for predicting total scores from partial scores using a test of 60 items selected from the Stanford Mental Arithmetic Test, Level III (Olshen, 1975). All students took all items. To make comparisons of predictions by each of the above methods with the observed total-test scores, we simulated an a posteriori matrix sampling design. Two sampling procedures were applied. First, the items in the total test were presented approximately in ascending order of difficulty based on previous data. By systematically sampling every sixth item, all items were assigned to six nonoverlapping item samples, three of type A and three of type B. Each item sample contained 10 items. Nine 20-item subtests were formed with each combination of one type-A item sample and one type-B item sample. The second sampling procedure randomly assigned subtests to students.

Approximately 450 students enrolled in Grades 3 through 5 were administered the test. Roughly 150 students therefore "took" each item sample and 50 "took" each subtest consisting of one type-A item sample and one type-B item sample. All subtests administered to examinees were overlapping, although within a type all item samples were nonoverlapping. The total-test score indicates the number of items that the examinee is predicted to have answered correctly.

Results

The observed total-test scores, O, were compared to the predicted total-test scores, E, for each method using the mean squared error, $\frac{\sum(O-E)^2}{n}$.

The results are shown in Table 1.

Insert Table 1 about here

Discussion

Three of these methods yielded fairly low mean squared errors, namely method 1 (regression with perfectly correlated nonoverlapping item samples), method 3 (regression with correlation between item samples on overlapping subtests), and method 4 (perfectly parallel overlapping or nonoverlapping item samples). Method 3 was slightly better than the other two. It is interesting to note that method 5, using item content with uncorrelated structural variable coefficients, also yields fairly good predictions.

Because the items were randomly assigned to item samples, we assume that the assumptions of methods 1 and 4 were not violated, i.e., the correlation between item samples was quite high and the item samples were approximately perfectly parallel. These correlations, shown in Table 2, are based only on those students who hypothetically took each of the nine pairs of item samples. Thus, method 3 is more robust and will accommodate violations of these assumptions, whereas methods 1 and 4 will not. For this reason, we recommend that a matrix sampling design intended to be used in estimating individual scores employ overlapping subtests and that method 3 be used to predict the total-test scores from partial scores.

Insert Table 2 about here

We further recommend that research be directed toward assessing the relative merits of the above three methods when the tests are not perfectly parallel and the correlation between subtests is less than unity. In addition, the methods of Bunda and Rasch may be compared to method 3 if large samples are obtained and the items conform to the Rasch model.

References

Bunda, M. A. An investigation of an extension of item sampling which yields individual scores. Journal of Educational Measurement, 1973, 10, 117-130.

Cook, D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Journal of Educational Measurement, 1967, 4, 27-33 (a).

Cook, D. L., & Stufflebeam, D. L. Estimating test norms from variable size item and examinee samples. Educational and Psychological Measurement, 1967, 27, 601-610 (b).

Jaeger, R. M. A comparison of three item-sampling procedures. Journal of Applied Psychology, in press.

Kleinke, D. J. A linear-prediction approach to developing test norms based on matrix-sampling. Educational and Psychological Measurement, 1972, 32, 75-84.

Lord, F. M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 32, 75-84.

Olshen, J. S. The use of performance models in establishing norms on a mental arithmetic test (Tech. Rep. 259). Stanford, Calif.: Institute for Mathematical Studies in the Social Sciences, 1975.

Owens, T. R., & Stufflebeam, D. L. An experimental comparison of item sampling and examinee sampling for estimating test norms. Journal of Educational Measurement, 1969, 6, 75-83.

Plumlee, L. B. Estimating means and standard deviations from partial data--an empirical check on Lord's item sampling technique. Educational and Psychological Measurement, 1964, 24, 623-630.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Shoemaker, D. M., & Knapp, T. R. A note on terminology and notation in matrix sampling. Journal of Educational Measurement, 1974, 11, 59-61.

Sirotnik, K. & Wellington, R. Incidence sampling: An integrated theory for "matrix sampling." Journal of Educational Measurement, 1977, 14, 343-399.

Suppes, P., & Morningstar, M. Computer-assisted instruction at Stanford, 1966-68. New York: Academic Press, 1972.

Wright, B. D., & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Table 1

Comparison of Methods

Method	Mean squared error
1. Perfectly correlated nonoverlapping item samples	13.30
2: Zero correlation between nonoverlapping item samples	111.93
3: Correlation between item samples on overhapping subtests	12.23
4: Perfectly parallel overlapping or nonoverlapping item samples	12.81
5: Item content with uncorrelated structural variable coefficients on overlapping or nonoverlapping subtests	15.79
6: Item content with perfectly correlated structural variable coefficients on overlapping or nonoverlapping subtests	35.78

Table 2

Means, Standard Deviations, and Correlation Matrix of Item Samples

Type-A item sample taken	Type-B item sample taken	Mean of type-A item sample	Mean of type-B item sample	Standard deviation of type-A item sample	Standard deviation of type-B item sample	Correlation between type-A and type-B item samples
1	1	5.84	6.08	1.54	1.91	.752
1	2	6.02	5.90	1.94	2.07	.779
1	3	5.98	5.51	1.75	2.07	.726
2	1	6.04	6.25	2.30	1.91	.784
2	2	5.86	6.14	2.11	2.00	.635
2	3	6.37	6.35	2.25	2.06	.764
3	1	6.27	6.31	2.43	2.10	.872
3	2	6.76	6.63	1.78	2.11	.721
3	3	6.43	5.47	1.72	1.67	.680

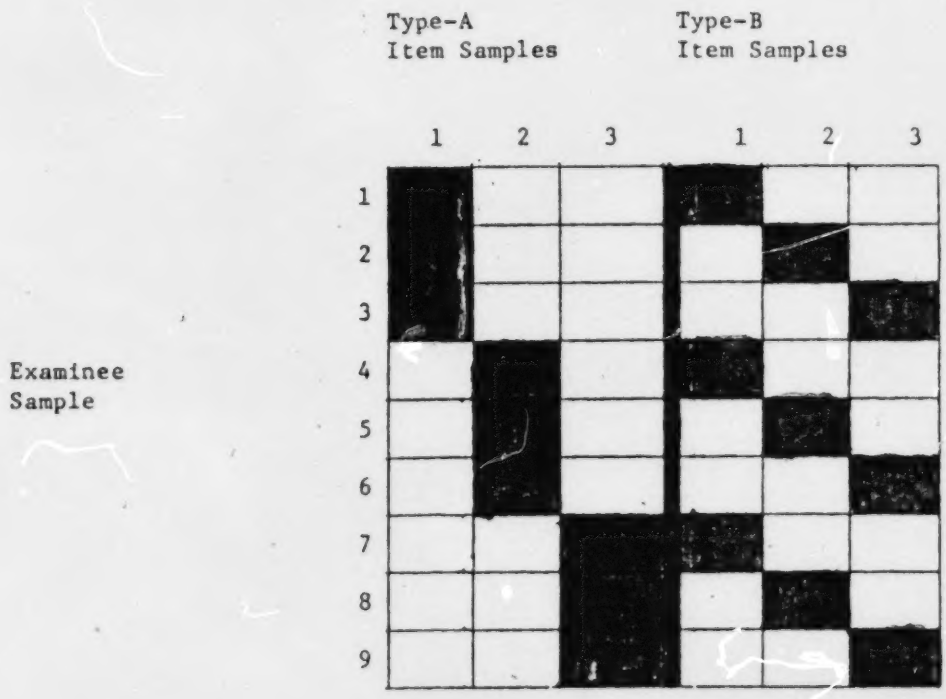

Figure 1. Example of matrix sampling design with 9 examinee samples.

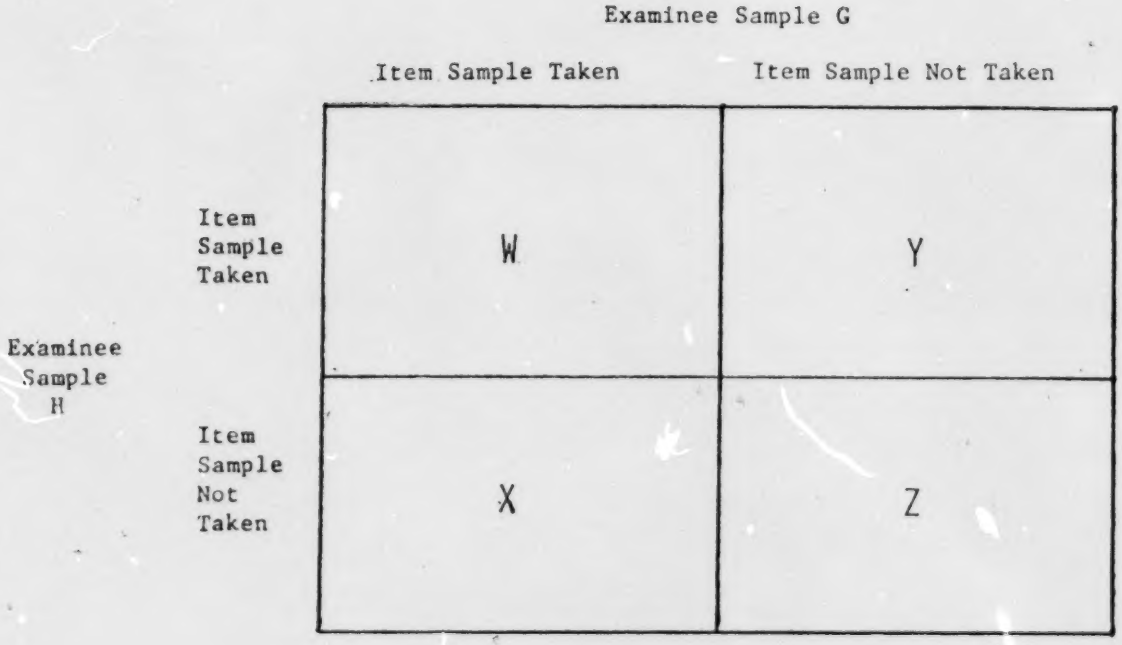

Figure 2. Item samples taken by two examinee samples.